心理学报 2020, Vol. 52, No. 12, 1452−1465
*Acta Psychologica Sinica*

https://dx.doi.org/10.3724/SP.J.1041.2020.01452

# 基于基尼指数的双目标 CD-CAT 选题策略*

罗 芬[1,2] 王晓庆[2] 蔡 艳[1] 涂冬波[1]

([1]江西师范大学心理学院, 南昌 330022) ([2]江西师范大学计算机信息工程学院, 南昌 330022)

**摘 要** 双目标 CD-CAT 的测验结果既可用于形成性评估也可用于终结性评估。基尼指数可度量随机变量的不确定性程度, 值越小则随机变量的不确定程度越低。本文用基尼指数度量被试知识状态类别以及能力估计置信区间后验概率的变化, 提出基于基尼指数的选题策略。Monte Carlo 实验表明与已有的选题策略相比, 新策略的知识状态分类精度和能力估计精度都较高, 同时能有效兼顾题库利用均匀性, 并能快速实时响应, 且受认知诊断模型和被试知识状态分布的影响较小, 可用于实际测验中含多种认知诊断模型的混合题库。

**关键词** 认知诊断, 项目反应理论, 基尼指数, 双目标 CD-CAT, 选题策略

**分类号** B841



## 1 引言

终结性评价用一个连续标量 $\theta$ (常称为潜在特质或能力)来刻画学生在某个学习阶段的学习效果, 基于项目反应理论(item response theory, IRT)的计算机化自适应测验(computerized adaptive testing, CAT)以“量体裁衣”的方式能更高效地实施终结性评估。形成性评价用一个离散向量 $\alpha$ (常称为潜在认知模式或知识状态)来帮助教师了解每个学生的潜在认知状态, 为教师提供教学反馈, 以便更好地“因材施教”, 这有利于学生学业和教师职业发展, 基于认知诊断理论(cognitive diagnostic theory, CDT)的 CAT 以“个性化”测验方式快速诊断被试认知的长处和短板。教学需要终结性评价与形成性评价相互结合, 既关注结果又关注过程, 使学习过程和对学习结果的评价达到和谐统一。IRT-CAT 关注终结性评价, CD-CAT (cognitive diagnostic computerized adaptive testing, CD-CAT)关注形成性评价, 两者结合的双目标 CD-CAT (dual objective CD-CAT, Dual-CAT)可以将它们的优势互补, 从而更好地完成测验目标。

Dual-CAT 的两个重点研究主题：一是建构题库的心理计量学指标, 正如 IRT-CAT 依赖于项目反应模型(item response method, IRM), CD-CAT 依赖于认知诊断模型(cognitive diagnostic model, CDM), Dual-CAT 也依赖于测验模型, 测验模型与题库的心理计量学指标息息相关。现有文献, 只有统一模型(unified model, 也称为 fusion model) (Hartz, 2002; Rupp et al., 2010)和高阶模型(de la Torre & Douglas, 2004)将被试的知识状态 $\alpha$ 与能力 $\theta$ 建构在一个模型中, 但统一模型所含参数较多, 在统计上难以估计(Hartz, 2002), 因此实际应用较少。而高阶模型采用层级结构, 将潜在特质视为比潜在属性更高层的一般能力, 能力 $\theta$ 与项目的正确作答概率之间的关系是通过被试知识状态 $\alpha$ 间接相关, 只有当属性个数较多时(例如大于 10), 能力 $\theta$ 的估计才会比较准确(de la Torre & Douglas, 2004; Hsu & Wang, 2015; Huang, 2020)。因此 Dual-CAT 的选题策略研究大多并不基于上述两种模型而采用分离建模的方法, 使用统一模型还是使用分离建模这两种方式决定了选题策略的构造方法也不同, 对于分离建模方式需要 IRM 和 CDM 的模型参数, 如何为这两套模型参数建立联系是实施 Dual-CAT 的基础。

de la Torre 和 Douglas (2004)的研究表明对于

收稿日期: 2019-10-14

* 国家自然科学基金(61967009, 31660278, 31760288, 31960186)和江西省教育厅科学技术研究项目(GJJ150356, GJJ160282)资助。

通信作者: 涂冬波, E-mail: tudongbo@aliyun.com

同一批数据, 高阶模型估计的 $\theta$ 与 IRT 中 2PLM (two-parameter logistic model)模型估计的 $\theta$ 有较高的相关性; Wang 等人(2014)的研究也表明, 单维项目反应模型(IRM)和 DINA 模型(Junker & Sijtsma, 2001)在属性间高度相关或线性层级相关时, 能够很好地拟合相同的数据, 他们的研究为分离建模方式提供了支持, 采用两步估计方法通过考虑各自的心理模型可获得稳定的 $\alpha$ 和 $\theta$ 估计(Kang et al., 2017)。

二是选题策略。选题策略是实施 Dual-CAT 的关键技术, 优良的选题策略应该既能达到较高的分类精度和估计精度以满足测验目的, 又能保证较为均匀的题库利用率以提高题库安全, 还需具有较快的运算速度以满足实时响应的需求, 研究者们围绕这个目标提出了多种选题策略。

IRT-CAT 和 CD-CAT 的选题策略分别注重潜在特质的评估和潜在认知结构的评估, 如何将这两者有效地结合起来? 学者们提出了若干种适合 Dual-CAT 的选题策略, 文献中已有的 Dual-CAT 选题策略主要有两类: 第一类是影子测验选题法; 第二类是组合策略选题法。

McGlohen 和 Chang (2008)在分离建模方式下讨论了影子测验选题法与 IRT-CAT 和 CD-CAT 的单一目标选题法的性能: (1)利用 IRT-CAT 中最大信息量策略(maximum fisher information, MFI) (Lord, 1980)或极大化 Kullback-Leibler (KL) (Chang & Ying, 1996)信息量策略选择适合被试当前估计能力 $\hat{\theta}$ 的项目, 测验结束再估计被试的知识状态 $\hat{\alpha}$; (2)利用 CD-CAT 中的极小化香农熵策略(Shannon entropy, SHE)或极大化 KL 信息量(Tatsuoka, 2002; Xu et al., 2003)选择适合被试当前知识状态估计值 $\hat{\alpha}$ 的项目, 测验结束再估计被试的能力 $\hat{\theta}$; (3)适应被试当前能力估计值 $\hat{\theta}$ 和知识状态估计值 $\hat{\alpha}$ 的影子测验(shadow test)选题, 即先根据被试能力估计值 $\hat{\theta}$, 采用(1)的方法构建最合适 $\hat{\theta}$ 的影子题库, 再从影子题库中采用(2)的方法选取最适合当前知识状态估计值 $\hat{\alpha}$ 的项目作为下一题的备选。他们将这三种方案在能力 $\theta$ 估计精度、认知状态 $\alpha$ 分类精度和项目曝光控制等 3 个指标上进行对比, 研究结果表明影子测验选题的表现更优。

杜宣宣(2010)也采用了影子测验选题法, 与 McGlohen 和 Chang (2008)不同之处在于, 他先构建最适合当前知识状态估计值 $\hat{\alpha}$ 的影子题库, 再从影子题库中选取最适合当前能力估计值 $\hat{\theta}$ 的项目作为下一题的备选, 并在不同属性层级结构下对能力 $\theta$ 估计精度、知识状态 $\alpha$ 分类精度等指标进行对比, 他的研究结果也表明与单一目标选题策略相比, 影子测验选题的表现更优。

McGlohen 和 Chang (2008)、杜宣宣(2010)的影子测验选题是两步估计法, 有学者认为(Cheng, 2007; Dai et al., 2016)两步“局部优化”的组合并不一定保证“良好的综合结果”, 更理想的项目选择方法应该在一个步骤内同时考虑 $\hat{\alpha}$ 和 $\hat{\theta}$ 以获得更适合的项目, 因此提出基于 $\hat{\alpha}$ 和 $\hat{\theta}$ 的组合策略选题法。

Cheng (2007)和 Dai 等人(2016)用线性加权组合 $objective = w * f(\hat{\theta}) + (1-w) * g(\hat{\alpha})$ 的指标代替影子测验选题, $f(\hat{\theta})$ 是关于 $\hat{\theta}$ 的信息量, 如 MFI 或 KL 等, $g(\hat{\alpha})$ 是关于 $\hat{\alpha}$ 的信息量, 如 SHE、KL、PWKL (posterior-weighted KL) (Cheng, 2009)、MPWKL (modified PWKL) (Kaplan et al., 2015)和 PWACDI (posterior-weighted attribute cognitive discrimination index) (Zheng & Chang, 2016)等。他们的研究表明在能力 $\theta$ 估计精度、认知状态 $\alpha$ 分类精度和项目曝光控制等 3 个指标上, 与影子测验选题法相比, 合成指标表现更优。

Wang 等人(2012)也基于分离建模方式, 将对被试知识状态的诊断视为约束条件, 使用 IRT-CAT 中最大优先级指标方法(maximum priority index, MPI) (Cheng & Chang, 2009)来选题, 即一种乘法组合策略: $objective = MPI_j(\hat{\alpha}) * f(\hat{\theta})$, 使得 IRT-CAT 既可以测量被试能力又能对被试认知状态进行分类。他们的研究表明, 由 KL 信息量构造的 MPI 指标能够获得较好的测量精度。

综合来看, 组合策略相对于影子测验选题法而言, 能更加细致地刻画 $\hat{\alpha}$ 和 $\hat{\theta}$ 之间相互作用对选题的影响。究竟采用加法组合策略还是乘法组合策略, 与 $f(\hat{\theta})$ 和 $g(\hat{\alpha})$ ($MPI_j(\hat{\alpha})$)采用何种信息量度量有关。Zheng 等人(2018)对比了多种信息量的加法组合策略和乘法组合策略, 他们的研究结果表明这两种组合方式在不同信息量下各有优劣。

加法组合策略的研究有 Cheng (2007)的两种 KL 信息量组合的 DIM (dual information method)策略, Wang 等人(2014)为消除 KL 和 PWKL 信息量差异提出的 ASI (aggregate standardized information method)策略和 ARI (aggregate ranked information method)策略, Kang 等人(2017)用对称 KL 信息量提出的 JSD (Jensen shannon divergence)策略以及 KL 和 MPWKL 信息量组合的 MASI (modified ARI)和

MARI (modified ASI)等。

乘法组合策略的研究有 Wang 等人(2012)提出的MPI的加权策略, Dai等人(2016)用对数转换消除MFI信息量和SHE信息量量纲差异, 将加法组合策略转换为乘法组合策略的 DWI (dapperness with information)策略, Zheng 等人(2018)提出的 IPA (information product approach)策略等。

这些选题策略在一定条件下, 都有各自的优势, 或精度较高但因运算量大选题耗时较多, 如 IPA 策略; 或精度稍低但可预先计算减少选题用时, 如 ASI 策略; 或精度更低但用时少且题库利用率较均匀, 如 JSD 策略。另外这些选题策略, 还可能存在因两种信息量量纲差异较大造成合成指标有所偏向, 或因进行转换以消除量纲差异所带来的信息损失等问题。我们希望开发一种对 $\hat{\alpha}$ 和 $\hat{\theta}$ 而言量纲比较统一的信息指标, 既保证估计精度和分类精度较高, 又能兼顾题库利用率均匀性且选题耗时较少的新策略。

在 CD-CAT 中, 大多采用贝叶斯决策对被试进行分类, 被试的知识状态类别是一个随机变量, 当类条件概率和先验概率已知的情况下, 通过贝叶斯公式计算被试属于每个类别的后验概率, 将被试的类别决策为后验概率大的一类, 理论上已证明这种决策的平均错误率最低(张学工, 2010, pp.14−15), 因此贝叶斯决策通常也称最小错误率贝叶斯决策。研究表明(陈平 等, 2011; 韩雨婷 等, 2018; Wang & Chang, 2011), 基于被试知识状态类别的后验概率所构造的选题策略(Zheng & Chang, 2016)和基于被试能力估计置信区间的后验概率所构造的选题策略具有较高分类精度和估计精度, 如 CD-CAT 中的香农熵策略(Tatsuoka, 2002; Xu et al., 2003)和多维 IRT-CAT 中连续熵(也称微分熵)策略(Wang & Chang, 2011; 韩雨婷 等, 2018)。

熵用于度量随机变量不确定性, 熵越大, 随机变量的不确定性就越大。在 CD-CAT 中, 用熵度量被试知识状态类别后验概率的变化, 然后采用贝叶斯决策根据被试知识状态类别的后验概率进行分类, 熵的变化直接反映各类别后验概率的变化, 因而基于熵所构建的选题策略的分类准确性较高, 如香农熵策略(Tatsuoka, 2002; Xu et al., 2003)。统计学中, 基尼指数也是一种度量随机变量不确定性的指标, 并应用于决策树的分类算法, 如既有基于熵的 ID3 算法(Quinlan, 1986)和 C4.5 算法(Quinlan, 1993), 也有基于基尼指数的 CART 算法(Breiman et al., 1984), 这些算法都是机器学习中的经典算法(周志华, 2016)。

本研究拟采用基尼指数构建双目标 CD-CAT 的选题策略。基尼指数和熵有共性也有差异。两者的共性在于它们都可以度量随机变量的不确定性程度且既可以处理连续型随机变量又可以处理离散型随机变量。设离散型随机变量 $X$ 所有可能取的值为 $x_v(v=1,2,\cdots,V)$, $X$ 取各个可能值的概率 $p\{X=x_v\}=p_v, v=1,2,\cdots,V$, 且 $\sum_{v=1}^{V}p_v=1$, 那么随机变量 $X$ 的熵可以表示为: $Ent(X)=-\sum_{v=1}^{V}p_v\ln p_v$, 随机变量 $X$ 的基尼指数可以表示为: $Gini(X)=\sum_{v=1}^{V}p_v(1-p_v)=1-\sum_{v=1}^{V}p_v^2$。令 $f(y)=-\ln y$, 在 $y$=1 处进行一阶泰勒展开(忽略高阶无穷小), $f(y)=f(1)+f'(1)(y-1)+O(\bullet)\approx 1-y$, 因此, 在 $p_v=1$ 处熵可近似转化为: $Ent(X)=-\sum_{v=1}^{V}p_v\ln p_v=\sum_{v=1}^{V}p_v(-\ln p_v)\approx\sum_{v=1}^{V}p_v(1-p_v)=Gini(X)$ [1], 说明在极值点处, 信息熵和基尼指数取得相同值。从数学表达式上看, 熵对随机变量的概率使用对数加权, 反映的是一种非线性关系, 而基尼指数使用线性加权, 反映的是一种线性关系。熵的计算公式中含有对数运算, 基尼指数只需求平方和, 因此基于基尼指数构造的选题策略会和香农熵选题策略一样具有较高的分类精度, 而运算速度快于香农熵策略, 且基尼指数的线性加权方式对测验过程中各类别的后验概率变化更加敏感, 从而有助于扩大选题范围, 有利于提高题库利用率。

本文利用基尼指数的上述优良性质, 提出基于基尼指数的选题策略, 期望新策略能保证测量精度, 同时兼顾题库利用均匀性并能快速实时响应, 为同时兼顾宏观能力评估和微观认知诊断提供新的更优的方法。

## 2 已有双目标 CD-CAT 选题策略简述

我们介绍三种有代表性的 Dual-CAT 的选题策略。ASI 策略是加法组合策略的代表, 通过标准化消除了两种信息量量纲差异后再将转换后的信息量进行线性加权; IPA 策略是乘法组合策略的代表;

[1] 摘自 https://www.jianshu.com/p/75518e6a5c64

JSD 策略是题库利用率最均匀且选题耗时最少的选题策略代表。

### 2.1　ASI 策略

Cheng (2009)提出用 PWKL 策略代替 KL 策略, 极大地提高了被试的知识状态 $\alpha$ ($\alpha$ 是一个 0 和 1 构成的向量)的分类精度, 设测验测量 $K$ 个独立属性, 被试的知识状态有 $2^K$ 类, 测验结束将被试划分到其中的一类, PWKL 选题策略的目标函数为:

$$Objective = \arg\max_{j \in R_t}(PWKL_j(\hat{\alpha})) \quad (1)$$

$$PWKL_j(\hat{\alpha}) = \sum_{c=1}^{2^K}[\pi_t(\alpha_c \mid Y) * KL_j(\hat{\alpha} \parallel \alpha_c)] \quad (2)$$

$$KL_j(\hat{\alpha} \parallel \alpha) = \sum_{y=0}^{1}\log\left(\frac{p(Y_j = y \mid \hat{\alpha})}{p(Y_j = y \mid \alpha)}\right) * p(Y_j = y \mid \hat{\alpha}) \quad (3)$$

其中 $R_t$ 为被试作答 $t$ 题后的剩余题库。$j$ 为剩余题库中的项目, $c = 1, 2, \ldots 2^K$ 为被试知识状态的类别下标, $\alpha_c$ 为 $2^K$ 种知识状态的第 $c$ 个类别, $\pi_t(\alpha_c \mid Y)$ 为在 $t$ 个项目的得分模式 $Y = (Y_1, Y_2, \ldots, Y_t)$ 下类别 $\alpha_c$ 的后验概率, $Y_j$ 为被试在项目 $j$ 的得分, $y$ 为项目的可能得分, 对于两级评分项目而言, $y$=0 或 1, $\hat{\alpha}$ 为被试知识状态的当前估计值, $p(Y_j = y \mid \hat{\alpha})$ 为给定 CDM 和已知 $\hat{\alpha}$ 时, 被试作答第 $j$ 题的答对概率。

Chang 和 Ying (1996)用 KL 策略代替 MFI 策略来测量被试的能力 $\theta$ ($\theta$ 是一个连续变量), 以克服当作答项目比较少时能力估计不准确的问题, KL 选题策略的目标函数为:

$$Objective = \arg\max_{j \in R_t}(KL_j(\hat{\theta})) \quad (4)$$

$$KL_j(\hat{\theta}) = \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} K_j(\hat{\theta} \parallel \theta)d\theta \quad (5)$$

$$K_j(\hat{\theta} \parallel \theta) = \sum_{y=0}^{1}\log\left(\frac{p(Y_j = y \mid \hat{\theta})}{p(Y_j = y \mid \theta)}\right) * p(Y_j = y \mid \hat{\theta}) \quad (6)$$

其中 $\delta$ 建议取 $3/\sqrt{t}$, $t$ 为被试已作答的项目数, $\hat{\theta}$ 为能力 $\theta$ 的当前估计值, $p(Y_j = y \mid \hat{\theta})$ 为给定 IRT 中的 IRM 和已知 $\hat{\theta}$ 时, 被试作答第 $j$ 题的答对概率。

Cheng (2007)提出 DIM 选题策略, 将关于 $\hat{\theta}$ 的 KL 信息和关于 $\hat{\alpha}$ 的 KL 信息线性组合为单个信息量以满足双目标 CD-CAT 选题的要求, DIM 选题策略的目标函数为:

$$Objective = \arg\max_{j \in R_t}(DIM_j(\hat{\alpha}, \hat{\theta})) \quad (7)$$

$$DIM_j(\hat{\alpha}, \hat{\theta}) = w * KL_j(\hat{\alpha}) + (1-w) * KL_j(\hat{\theta}) \quad (8)$$

其中 $w$ 为权重。

Wang 等人(2014)将 DIM 策略中关于 $\hat{\alpha}$ 的信息度量用 PWKL 信息量代换, 并认为 $PWKL_j(\hat{\alpha})$ 和 $KL_j(\hat{\theta})$ 量纲不一致, 可采用标准化方法消除两者之间的差异, 进而提出了 ASI 策略, ASI 选题策略的目标函数为:

$$Objective = \arg\max_{j \in R_t}(ASI_j(\hat{\alpha}, \hat{\theta})) \quad (9)$$

$$ASI_j(\hat{\alpha}, \hat{\theta}) = w * PWKL_j^*(\hat{\alpha}) + (1-w) * KL_j^*(\hat{\theta}) \quad (10)$$

$$PWKL_j^*(\hat{\alpha}) = \frac{(PWKL_j(\hat{\alpha}) - mean(PWKL(\hat{\alpha})))}{SD(PWKL(\hat{\alpha}))} \quad (11)$$

$$KL_j^*(\hat{\theta}) = \frac{(KL_j(\hat{\theta}) - mean(KL(\hat{\theta})))}{SD(KL(\hat{\theta}))} \quad (12)$$

其中 $mean(PWKL(\hat{\alpha}))$ 为剩余题库 $R_t$ 所有项目关于被试知识状态当前估计值 $\hat{\alpha}$ 的 PWKL 信息量均值, $SD(PWKL(\hat{\alpha}))$ 为其标准差。$mean(KL(\hat{\theta}))$ 为剩余题库 $R_t$ 所有项目关于能力当前估计值 $\hat{\theta}$ 下 KL 信息量的均值, $SD(KL(\hat{\theta}))$ 为其标准差。Wang 等人(2014)还建议权重 $w$ 取值为 $1 - t / TL$, $t$ 为已做答项目数, $TL$ 为预设的测验长度。

### 2.2　IPA 策略

Zheng 和 Chang (2016)提出适用于 CD-CAT 短测验的 PWACDI (posterior-weighted attribute cognitive discrimination index)选题策略, PWACDI 选题策略的目标函数为:

$$Objective = \arg\max_{j \in R_t}(PWACDI_j) \quad (13)$$

$$PWACDI_j = \sum_{k=1}^{K}\frac{1}{2^K}\sum_{all\ relevant\ cells} PWD_{juv} \quad (14)$$

$$PWD_{juv} = \pi(\alpha_u \mid Y) * \pi(\alpha_v \mid Y) * \sum_{y=0}^{1}\left[p(Y_j = y \mid \alpha_u) * \log\left(\frac{p(Y_j = y \mid \alpha_u)}{p(Y_j = y \mid \alpha_v)}\right)\right] \quad (15)$$

其中, $u$ 和 $v$ 为被试知识状态的类别下标, $\alpha_u$ 和 $\alpha_v$ 为 $2^K$ 种知识状态中不相同的两个类别, $PWD_{juv}$ 为根据项目 $j$ 构造的 $2^K \times 2^K$ 的 KL 信息矩阵, 矩阵内的元素为任意两个知识状态的期望加权 KL 距离。all relevant cells 是指 $PWD_{juv}$ 矩阵中两种不同知识状态 $\alpha_u$ 和 $\alpha_v$ 所对应位置的所有元素, 且这两种知识状态仅在第 $k$ 个属性值是不同的, 其他属性值相同。PWACDI 选题策略与被试当前知识状态估计值 $\hat{\alpha}$ 无关, 并且注重区分 $2^K$ 种模式中, 那些差异较小的模式, 这不同于 PWKL 策略。

Zheng 等人(2018)提出适用于双目标 CD-CAT 的 IPA 策略, 认为该策略能提供一个统一的框架来

连接其他的双目标选题策略，将“权重”视为与 IRT 信息相等的对应项，则不需考虑公式(10)中的权重。信息量乘法的选题策略的目标函数为：

$$Objective = \arg\max_{j \in R_t}(P_j \times KL_j(\hat{\theta})) \tag{16}$$

$P_j$ 可以是 $PWKL_j(\hat{\alpha})$ 或 $PWACDI_j$ 等其他 CD-CAT 的选题策略，根据 Zheng 等人(2018)的研究，$PWACDI_j \times KL_j(\hat{\theta})$ 的表现更好。

### 2.3 JSD 策略

Kang 等人(2017)提出 JSD 选题策略，不同于 PWKL 策略，它是对称的 KL 信息，令 $\eta$=($\hat{\alpha}$,$\hat{\theta}$), JSD 选题策略的目标函数为：

$$objective = \arg\max_{j \in R_t}(JS_j(\hat{\alpha} \| \hat{\theta})) \tag{17}$$

$$JS_j(\hat{\alpha} \| \hat{\theta}) = \mathrm{w}^*\mathrm{KL}_j(\hat{\alpha} \| \eta) + (1-\mathrm{w})^*\mathrm{KL}_j(\hat{\alpha} \| \eta) \tag{18}$$

$$g_j(\eta) = w^* p(Y_j = y \mid \hat{\alpha}) + (1-w)^* p(Y_j = y \mid \hat{\theta}) \tag{19}$$

$$KL_j(\hat{\alpha} \| \eta) = \sum_{y=0}^{1} p(Y_j = y \mid \hat{\alpha})^* \log\left(\frac{p(Y_j = y \mid \hat{\alpha})}{g_j(\eta)}\right) \tag{20}$$

$$KL_j(\hat{\theta} \| \eta) = \sum_{y=0}^{1} p(Y_j = y \mid \hat{\theta})^* \log\left(\frac{p(Y_j = y \mid \hat{\theta})}{g_j(\eta)}\right) \tag{21}$$

特别说明，为了更清楚的描述 JSD 策略，我们补充了一些符号，因此本文中 JSD 选题策略中的表达式与原文(Kang et al., 2017)不是完全相同，但没有改变选题策略本身的含义。

## 3 基于基尼指数的双目标 CD-CAT 选题策略

本研究分别定义了基于被试知识状态类别的后验概率和基于被试能力估计置信区间的后验概率的基尼指数，并将两者组合构成基于基尼指数的双目标 CD-CAT 新策略，以期达成高精度、高题库利用率和快速反馈的测验需求。

### 3.1 基于基尼指数的 CD-CAT 选题策略

设测验考查 $K$ 个属性，在 $t$ 个项目的得分模式 $Y = (Y_1, Y_2, \cdots, Y_t)$ 下类别 $\alpha_c (c = 1, 2, \cdots, 2^K)$ 的后验概率为 $\pi_t(\alpha_c \mid Y)$ (简记为 $\pi_t(\alpha_c)$ )且 $\sum_{c=1}^{2^K} \pi_t(\alpha_c) = 1$，根据基尼指数的定义(李航，2012)，则被试知识状态类别后验概率的基尼指数定义为：

$$Gini_CD(\pi_t) = \sum_{c=1}^{2^K}[\pi_t(\alpha_c)^*(1-\pi_t(\alpha_c))] = 1 - \sum_{c=1}^{2^K}[\pi_t(\alpha_c)]^2 \tag{22}$$

$$\pi_t(\alpha_c) \propto \pi_0(\alpha_c)^* \prod_{h=1}^{t}[(p_h(\alpha_c))^{Y_h}(1-p_h(\alpha_c))^{1-Y_h}] \tag{23}$$

$\pi_t$ 为 $t$ 个项目的反应模式 $Y = (Y_1, Y_2, \cdots, Y_t)$ 下知识状态类别后验概率的集合，$\pi_0(\alpha_c)$ 是类别 $\alpha_c$ 的先验概率，一般取 $1/2^K$，$p_h(\alpha_c)$ 为给定 CDM 下知识状态为 $\alpha_c$ 的被试答对第 $h$ 题的概率，$Y_h$ 为被试在项目 $h$ 的得分，其他符号的含义同第 2 节。

$Gini_CD(\pi_t)$ 刻画在 $t$ 个项目的反应模式 $Y = (Y_1, Y_2, \cdots, Y_t)$ 下，被试知识状态类别后验概率分布的离散程度，其值越小则概率分布越集中，即一个或某些类别的后验概率会远大于其他类别，从而有助于提高贝叶斯决策对被试分类的准确性。遍历并选择剩余题库中使 $Gini_CD(\pi_t, Y_j)$ 取得最小值的项目 $j$ 作为下一题的候选。

由于被试对候选项目 $j$ 的作答反应 $Y_j$ 未知，对于两级评分项目，$Y_j$ 的值为 0 或 1 (即 $y = 0$ 或 1)，定义被试知识状态类别后验概率的期望基尼指数：

$$E[Gini_CD(\pi_t, Y_j)] = \sum_{y=0}^{1} Gini_CD(\pi_{t+1} \mid Y_j = y)^* P(Y_j = y \mid \pi_t) \tag{24}$$

由全概率公式

$$P(Y_j = y \mid \pi_t) = \sum_{c=1}^{2^K}[(p_j(\alpha_c))^y(1-p_j(\alpha_c))^{1-y}]\pi_t(\alpha_c) \tag{25}$$

$Gini_CD$ 选题策略的目标函数为：

$$Objective = \arg\min_{j \in R_t}(E[Gini_CD(\pi_t, Y_j)]) \tag{26}$$

$R_t$ 为被试的剩余题库，即从剩余题库中选择具有最小 $E[Gini_CD(\pi_t, Y_j)]$ 的项目 $j$。

### 3.2 基于基尼指数的 IRT-CAT 选题策略

在 IRT-CAT 测验初始阶段，由于被试当前能力估计值 $\hat{\theta}$ 往往与被试真实能力值偏差较大，此时基于 $\hat{\theta}$ 的 Fisher 信息量不是一个好的测验效率指示量，因此在测验初始阶段不能发挥重要作用(Chang & Ying, 1996)。Veerkamp 和 Berger (1994)提出用基于置信区间中信息函数的最高均值代替基于某一点的项目的区间信息选题准则，较好地克服了由于 $\hat{\theta}$ 估计不准带来的低效选题问题。

优良的选题策略使得被试能力估计值 $\hat{\theta}$ 随着测验的进行，越来越接近其真实值，根据 Chang 和 Ying (1996)以及 Wang 和 Chang (2011)中 KL 全局信息量和连续熵的定义，我们定义了基于被试能力估计值 $\hat{\theta}$ 的置信区间后验概率的基尼指数，它类似于 KL 全局信息量，利用区间信息代替某个估计点的信息。令 $\hat{\hat{\theta}}$=$\hat{\theta}$+i$\Delta\theta$，

$$Gini_IRT(\pi_t(\hat{\theta})) = \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta}[\pi_t(\theta)*(1-\pi_t(\theta))]d\theta \approx \sum_{i=-s}^{s}[\pi_t(\hat{\hat{\theta}})*(1-\pi_t(\hat{\hat{\theta}}))]\Delta\theta \tag{27}$$

$$\pi_t(\hat{\theta}) \propto \pi_0(\hat{\theta})*\prod_{h=1}^{t}[(p_h(\hat{\theta}))^{Y_h}(1-p_h(\hat{\theta}))^{1-Y_h}] \tag{28}$$

其中, $\pi_t(\hat{\theta})$ 为 $t$ 个项目的反应模式 $Y=(Y_1, Y_2,\cdots,Y_t)$ 下, 能力估计值 $\hat{\theta}$ 的置信区间内后验概率的集合, Chang 和 Ying (1996)建议 $\delta = 3/\sqrt{t}$, 根据 BILOG 程序中计算后验期望概率的推荐值, 取求积结点数为与 $2\sqrt{t}$ 相近的自然数, $s = \lceil 2\sqrt{t}\rceil/2$, "$\lceil\ \rceil$"表示向上取整, $\pi_0(\hat{\theta})$ 是 $\hat{\theta}$ 的先验概率, 若能力先验信息未知则取均匀分布。$p_h(\hat{\theta})$ 为给定 IRM 下能力为 $\hat{\theta}$ 的被试答对第 $h$ 题的概率, 其他符号的含义同第 2 节。

遍历并选择剩余题库中使 $Gini_IRT(\pi_t(\hat{\theta}),Y_j)$ 取得最小值的项目 $j$ 作为下一题的候选。

由于被试对候选项目 $j$ 的作答反应 $Y_j$ 未知, 对于两级评分项目, $Y_j$ 的值为 0 或 1(即 $y$=0 或 1), 定义能力估计值 $\hat{\theta}$ 的置信区间后验概率的期望基尼系数:

$$E[Gini_IRT(\pi_t(\hat{\theta}),Y_j)] = \sum_{y=0}^{1} Gini_IRT(\pi_{t+1}(\hat{\theta})\,|\,Y_j=y)*P(Y_j=y\,|\,\pi_t(\hat{\theta})) \tag{29}$$

$$P(Y_j=y\,|\,\pi_t(\hat{\theta})) = \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta}[(p_j(\theta))^y(1-p_j(\theta))^{1-y}]\pi_t(\theta)d(\theta) \tag{30}$$

$Gini_IRT$ 选题策略的目标函数为:

$$Objective = \mathop{\arg\min}_{j\in R_t}(E[Gini_IRT(\pi_t(\hat{\theta}),Y_j)]) \tag{31}$$

$R_t$ 为被试的剩余题库, 即从剩余题库中选择具有最小 $E[Gini_IRT(\pi_t(\hat{\theta}),Y_j)]$ 的项目 $j$。

### 3.3　组合策略

Cheng (2007)和 Wang 等人(2014)提出将基于被试知识状态 $\hat{\alpha}$ 的 KL 信息函数和能力 $\hat{\theta}$ 的 KL 信息函数进行加权线性组合以得到单一信息量形式的双目标选题策略, 如公式(8)和(10)。Zheng 等人(2018)提出将两个函数相乘的双目标选题策略, 如公式(16)。由于乘法运算更加费时。我们采用 Cheng (2007)和 Wang 等人(2014)的线性加权和方式获得基于基尼指数的双目标选题策略目标函数。

本文提出的新策略基于两个随机变量后验概率的基尼指数构造的新指标, 由于每个随机变量后验概率的取值范围为[0, 1], 且后验概率的累加和为 1, 这两个后验概率构造的基尼指数指标的量纲不会有很大差异, 不需要像 Wang 等人(2014)将两个 KL 信息量进行标准化再进行线性组合, 因转化还是会带来信息损耗, 新策略的合成方法保持了原有信息。

Gini 选题策略的目标函数为:

$$Gini_j(\hat{\alpha},\hat{\theta}) = w*E[Gini_CD(\pi_t,Y_j)] + (1-w)*E[Gini_IRT(\pi_t(\hat{\theta}),Y_j)] \tag{32}$$

$$Objective = \mathop{\arg\min}_{j\in R_t}(Gini_j(\hat{\alpha},\hat{\theta})) \tag{33}$$

其中, $w$ 是权重, 根据 Wang 等人(2014)的建议, 在高质量题库中建议使用理论权重 $w = 1 - t/TL$, $t$ 为已做答项目数, $TL$ 为预设的测验长度。

## 4　模拟实验设计

为考察不同 CDM、被试不同知识状态分布以及不同测验长度下新策略的性能及其与其他选题策略的比较, 开展了 Monte Carlo 模拟实验研究。实验考察了 3 种 CDM (G-DINA, DINA, R-RUM)× 3 种被试知识状态的分布(高阶模型、高相关多元正态模型和低相关多元正态模型) × 4 种测验长度(5、10、15、20) = 36 种情形下新策略的表现。

### 4.1　认知诊断模型

在饱和模型 G-DINA (de la Torre, 2011)和缩减模型(DINA, R-RUM) (Hartz, 2002; Junker & Sijtsma, 2001)下讨论各选题策略表现。G-DINA 模型在适当约束条件下可简化为不同的缩减模型: 若 G-DINA 所有主效应和低阶交互效应值为 0, 则其简化为 DINA 模型; 若对数连接函数的所有交互效应的值为 0, 则可得 R-RUM。

### 4.2　题库参数和被试知识状态

#### 4.2.1　模拟题库项目的属性向量

设题库考察 5 个独立属性, 每个项目最多考察 3 个属性即共 25 ($C_5^1+C_5^2+C_5^3=25$)种项目属性向量, 每种属性向量重复 10 次, 可得题库中 250 个项目的属性向量。

#### 4.2.2　模拟被试知识状态的真值

被试知识状态采用两种方式模拟, 一种采用 HO-CDM (Wang et al., 2012, 2014; Huang, 2020), 另一种采用多元正态分布生成(Dai et al., 2016; Kang et al., 2017)。考察这两种模拟方式是因为他们的作答反应数据可以同时拟合 CDM 和 IRT 的模型, 也是双目标 CD-CAT 中常用的模拟方法。

(1)被试知识状态用 HO-CDM (de la Torre &

Douglas, 2004)生成。高阶模型假定考生是否掌握某个属性与泛化的潜在能力有关。通过 logit 链接，给定高阶能力 $\theta_i$，被试 $i$ 掌握属性 $k$ 的概率定义为：$P(\alpha_{ik} \mid \theta_i) = \dfrac{\exp(\lambda_{1k}(\theta_i - \lambda_{0k}))}{1+\exp(\lambda_{1k}(\theta_i - \lambda_{0k}))}$，类似 IRT 中的 2PLM 模型，其中 $\lambda_{1k}$ 和 $\lambda_{0k}$ 是区分度参数和位置参数，$\theta_i \sim N(0,1)$，$\ln \lambda_{1k} \sim N(0,1)$ (将值的约束在[0.2, 2.5]区间范围内)，$\lambda_{0k} \sim N(0,1)$，另生成随机数 $r$，$r \sim uniform(0,1)$，如果 $P(\alpha_{ik} \mid \theta_i) \geqslant r$，则令 $\alpha_{ik}=1$，否则令 $\alpha_{ik}=0$ (Ma & de la Torre, 2020)。

(2)被试知识状态用多元正态模型生成。采用多元正态阈值模型(均值为 0；变量间的相关分别设 0.8，0.2 两种水平，分别代表属性间存在高相关和低相关)生成被试真实属性掌握模式，用 0 作为截断点获得离散值知识状态(Ma & de la Torre, 2020)。

#### 4.2.3 模拟题库 CDM 项目参数和 IRT 模型参数

采用第 1 节介绍的分离建模方法构建题库，CDM 模型分别采用 G-DINA、DINA 和 R-RUM 模型，IRT 模型采用 2PLM，这些模型是研究和实践中经常使用的模型。

题库参数用 R 软件中的 GDINA 包和 mirt 包模拟和估计。

以 G-DINA 模型和被试的知识状态采用高相关多元正态模型生成为例介绍题库项目参数的模拟。

(1)根据 GDINA 包(Ma & de la Torre, 2020)的说明文档，CDM 参数的设定可以采用三种方法。第一种方法，为每个项目指定猜测参数 $p(0)$和失误参数 $1-p(1)$，其中，$p(0)$表示未掌握项目任何一个考察属性的被试正确作答概率，$p(1)$表示掌握了项目所有考察属性的被试正确作答概率，其他类型的被试作答概率从[$p(0)$, $p(1)$]中生成，需符合约束单调性原则，即掌握项目考察属性个数多的被试的正确作答概率大于掌握项目所考察属性个数少的被试的正确作答概率；第二种方法，为每个项目的每种知识状态指定答对概率；第三种方法，为每个项目指定 G-DINA 模型中的 delta 参数。

因第一种方法简单易操作，本研究采用第一种方法，利用 GDINA 包中的 simGDINA 函数模拟 G-DINA 模型的项目参数，设 $p(0) \sim uniform(0.05,0.25)$，$p(1) \sim uniform(0.75,0.95)$，其他掌握了项目所考察的部分属性的被试正确作答概率从[$p(0)$, $p(1)$]中生成，正确作答概率保证单调性。

(2)因为 2PLM 的项目参数估计需要 1000 以上样本才能获得较好的精度，本文利用高相关多元正态模型模拟 3000 个被试的知识状态，根据已知的每个项目属性向量和 G-DINA 模型的项目参数获得每个被试在每个项目上的正确作答概率 $p$，另外生成随机数 $r$, $r \sim uniform(0,1)$，如果 $p \geqslant r$，则令得分为 1，否则令得分为 0，即获得 3000×250 的完全得分阵(Wang et al., 2012, 2014)。将得分阵用 R 软件中的 mirt 包(Chalmers, 2012)中 mirt 函数拟合 2PLM 可得题库中 250 个项目的区分度和难度参数，用 R 软件中的 GDINA 包中 GDINA 函数对 G-DINA 模型参数进行校正，以获得更准确的参数。

按照上述方法，可以获得相应的 3(G-DINA, DINA, R-RUM) × 3(高阶模型、高相关多元正态模型和低相关多元正态模型) = 9 种题库的 CDM 的参数和 2PLM 参数。

#### 4.2.4 模拟被试能力的真值

被试对项目的反应是根据 CDM 模型模拟生成，模拟被试作答题库所有项目的反应数据，将反应数据用期望后验算法(Bock & Mislevy, 1982)估计被试的能力值作为其真值(Wang et al., 2012, 2014; Dai et al., 2016; Kang et al., 2017)。

### 4.3 选题策略

DIM 策略(Cheng, 2007)是首个将两个 KL 信息量进行线性组合的策略，ASI 策略将两个信息量标准化以消除两个信息量的量纲差异后再线性组合，根据 Wang 等人(2014)的研究结果，ASI 策略优于 DIM 策略。根据 Zheng 等人(2016, 2018)的研究结果，PWACDI 策略在短测验上的分类精度优于 PWKL 策略，PWACDI*KL 策略和 DWI 策略(Dai et al., 2016)都属于双信息量的乘法组合策略 IPA，研究(Zheng et al., 2016, 2018)表明，PWACDI*KL 在一簇 IPA 策略中表现更好。JSD 策略(Kang et al., 2017)基于被试当前知识状态估计值和能力估计值的对称 KL 信息选题，在选题过程中不需要积分运算，因此运算简单，选题速度很快，根据 Kang 等人(2017)的研究，JSD 策略与其他策略相比在选题用时和题库利用均匀性上有较大的优势。

本文将 Gini 策略与 ASI 策略(Wang et al., 2014)、IPA 中的代表 PWACDI*KL 策略 (Zheng et al., 2018), JSD (Kang et al., 2017)策略在 9 种题库下进行对比，从测量精度(包含知识状态分类精度和能力估计精度)、题库利用均匀性和选题用时等方面考查新策略的性能。

### 4.4 终止规则

实验均采用定长测验，定长测验设置了 4 个水

平：5、10、15 和 20 题。

### 4.5 评价指标

#### 4.5.1 知识状态分类精度指标

模式判准率是评价知识状态分类精度的指标，值越大，分类精度越高。

$$PMR = \frac{\sum_{i=1}^{N} I(\hat{\alpha}_i = \alpha_i)}{N}$$

其中 $I(\bullet)$ 表示当条件 $\bullet$ 为 TRUE 时，计数为 1，否则为 0。$N$ 为被试人数。$\hat{\alpha}_i$ 是被试知识状态的估计值，$\alpha_i$ 是被试知识状态的真值。

#### 4.5.2 能力估计精度指标

用 *Bias* 和 *RMSE* 作为能力估计精度的指标。值越小，参数返真性越高。

$$Bias = \frac{1}{N}\sum_{i=1}^{N} |\hat{\theta}_i - \theta_i|$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2}$$

其中 $\hat{\theta}_i$ 被试能力估计值，$\theta_i$ 被试能力真值。其他变量含义同上。

#### 4.5.3 题库使用均匀性指标

卡方值和测验重叠率是评价题库使用均匀性的重要指标，值越小，题库使用越均匀，利用率越高。

卡方值指标 $$\chi^2 = \frac{\sum_{j=1}^{L}(m_j / N - TL / L)^2}{TL / L}$$

测验重叠率 $$TOE = \frac{N \times \sum_{j=1}^{L}(m_j / N)^2}{(N-1)\times TL} - \frac{1}{N-1}$$

其中 $m_j$ 为项目 $j$ 的曝光次数，$L$ 为题库容量，$TL$ 设定的测验长度，其他变量含义同上。

#### 4.5.4 选题用时

$$TC = \frac{\sum_{i=1}^{N} T_i}{N}$$

其中，$T_i$ 为第 $i$ 个被试完成测验所需时间(单位：秒)。由于模拟研究的时间消耗主要在选题上，其他用时可忽略不计，因此 $TC$ 即为选题耗时。值越小，选题速度越快。

### 4.6 CAT 实施过程

整个 CAT 的程序，运行于 Python 3，硬件配置为 4 核处理器 Intel Core i5 1.9GHz，内存 8G。以 G-DINA 模型和高相关多元正态模型模拟被试知识状态的实验条件为例，说明 CAT 的实施过程。

(1)选择对应实验条件下在 R 环境中用 GDINA 包和 mirt 包构建的题库(细节参照第 4.2 节);

(2)采用高相关多元正态模型模拟被试的知识状态作为被试知识状态的真值，并模拟被试采用 G-DINA 模型作答题库所有题，用期望后验法估计其能力值作为被试能力真值(细节参照第 4.2 节);

(3)随机分配 3 题给被试作答，根据初始 3 题的反应，估计被试知识状态初值和能力初值;

(4)分别采用 Gini 策略, ASI 策略, IPA 策略, JSD 策略选题进入各自 CAT 的过程，被试每作答一个项目，采用最大后验法估计被试知识状态和采用期望后验法估计被试能力;

(5)重复(4)直到满足测验停止要求;

(6)测验结束后根据每种策略下的最终被试知识状态估计值和被试能力估计值计算第 4.5 节中的评价指标。

为消除随机效应，每次模拟 1000 个被试，每种实验条件重复 10 次，计算每种实验条件下各评价指标的平均值(见第 5 节的表格, *SD* 表示其标准差)。

## 5 实验结果

### 5.1 分类精度的比较

表 1 表明，Gini 策略和 IPA 策略的模式判准率远高于 ASI 策略和 JSD 策略，且整体而言 Gini 策略的模式判准率略高于 IPA 策略，这两种策略在不同实验条件下的模式判准率均超过 95%且标准差都较小，说明他们的分类结果稳定可靠，可适用于不同 CDM 的题库或多种 CDM 混合题库。

图 1 是各选题策略在不同测验长度上的表现，随测验长度的增加，各选题策略的模式判准率逐渐提高。Gini 策略和 IPA 策略的变化曲线非常相似，增长最快，始终保持最好的判准率。在短测验($TL <$ 15)中，Gini、IPA 和 ASI 策略的模式判准率很接近，在中长测验($TL$>15)后，ASI 策略的增长速度要低于前两者。与表 1 的结论相同，Gini 和 IPA 策略在不同实验条件下的变化曲线没有太大差异，因此他们在短测验和中长测验下均能获得较好的分类精度。

### 5.2 能力估计精度的比较

表 2 表明，除在 DINA 模型下属性间低相关的实验条件外，4 种策略对能力估计基本是无偏的。ASI 策略的估计偏差最小，其次是 Gini 策略。IPA 策略具有最小的能力估计均方差值，与之相比，Gini 策略稍稍差一些，但最大差异也仅有 0.04。当属性间高相关时，4 种选题策略的能力估计均方差值非常接近，最大差异仅有 0.03，而在其他条件下，最

**表 1　20 题各选题策略的模式判准率均值及标准差**

| CDM 模型 | 知识状态生成模型 | 选题策略 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Gini | | ASI | | IPA | | JSD | |
| | | *Mean*/% | *SD* | *Mean*/% | *SD* | *Mean*/% | *SD* | *Mean*/% | *SD* |
| G-DINA | HO | **97.00** | 0.009 | 89.28 | 0.025 | 96.10 | 0.010 | 85.04 | 0.024 |
| | MV-0.8 | 97.22 | 0.004 | 93.05 | 0.011 | **97.44** | 0.008 | 92.02 | 0.014 |
| | MV-0.2 | **96.84** | 0.007 | 90.78 | 0.014 | 96.35 | 0.006 | 87.51 | 0.016 |
| DINA | HO | **97.45** | 0.010 | 90.99 | 0.032 | 97.18 | 0.011 | 75.31 | 0.060 |
| | MV-0.8 | **97.24** | 0.011 | 93.45 | 0.017 | 97.06 | 0.010 | 91.46 | 0.023 |
| | MV-0.2 | **97.57** | 0.006 | 93.76 | 0.007 | 96.93 | 0.008 | 86.23 | 0.050 |
| R-RUM | HO | **95.41** | 0.010 | 87.61 | 0.021 | 95.38 | 0.010 | 76.64 | 0.028 |
| | MV-0.8 | **97.09** | 0.009 | 92.45 | 0.014 | 96.82 | 0.008 | 91.67 | 0.010 |
| | MV-0.2 | 96.81 | 0.008 | 87.88 | 0.022 | **96.82** | 0.012 | 80.52 | 0.038 |

注：HO 指被试知识状态用 HO-CDM 生成, MV-0.8 指被试知识状态用多元正态模型生成且属性间相关系数为 0.8, MV-0.2 指被试知识状态用多元正态模型生成且属性间相关系数为 0.2。

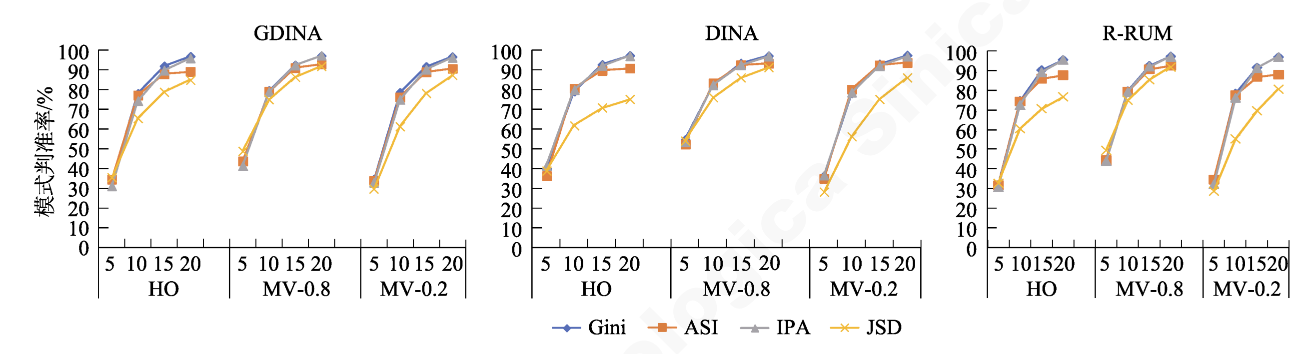



图 1　不同测验长度的模式判准率

**表 2　20 题各选题策略的 *Bias* 和 *RMSE***

| CDM 模型 | 知识状态生成模型 | 选题策略 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Gini | | ASI | | IPA | | JSD | |
| | | *Bias* | *RMSE* | *Bias* | *RMSE* | *Bias* | *RMSE* | *Bias* | *RMSE* |
| G-DINA | HO | 0.02 | 0.32 | 0.00 | 0.41 | 0.04 | **0.28** | 0.02 | 0.40 |
| | MV-0.8 | 0.00 | **0.29** | 0.01 | **0.29** | 0.02 | **0.29** | 0.02 | 0.30 |
| | MV-0.2 | 0.03 | **0.27** | 0.02 | 0.32 | 0.07 | **0.27** | 0.05 | 0.42 |
| DINA | HO | –0.08 | 0.40 | –0.02 | 0.41 | –0.14 | **0.37** | –0.05 | 0.46 |
| | MV-0.8 | 0.02 | 0.34 | 0.01 | **0.32** | –0.03 | 0.35 | –0.08 | 0.35 |
| | MV-0.2 | –0.12 | 0.38 | –0.09 | **0.36** | –0.24 | 0.42 | –0.28 | 0.52 |
| R-RUM | HO | –0.07 | **0.35** | –0.01 | 0.42 | –0.14 | **0.35** | –0.02 | 0.45 |
| | MV-0.8 | 0.00 | **0.30** | –0.02 | **0.30** | –0.03 | **0.30** | –0.03 | 0.32 |
| | MV-0.2 | –0.04 | 0.31 | –0.01 | 0.43 | –0.10 | **0.29** | –0.05 | 0.51 |

大差异达 0.22, 这说明属性间高相关时, 4 种选题策略均可用, 而其他条件下可优先考虑 IPA 和 Gini 策略。Gini 和 IPA 策略的能力估计精度与 CDM 有关, Gini 策略所受影响更小一些。ASI 和 JSD 策略的能力估计精度既与 CDM 有关又与被试知识状态分布有关。

图 2 表明随测验长度的增加被试能力估计的均方差值在下降, 即参数估计精度在上升, Gini 和 IPA 策略均方差值下降速度最快, 且两种策略的下降曲线基本相同, JSD 策略的下降趋势最慢。当属性间高相关时, 4 种选题策略的曲线基本重合, 在其他条件下, 与图 1 类似, 在短测验($TL < 15$)中, Gini、IPA 和 ASI 策略的曲线基本一致, 在中长测验($TL > 15$)后, ASI 策略不如前两者。因此 Gini 和

IPA 策略在短测验和中长测验下均能获得较好的能力估计精度。

### 5.3　题库使用均匀性的比较

表 3 表明, JSD 策略的题库利用均匀性优于其他 3 种策略。Gini 和 IPA 策略的题库利用率指标值相近, 整体而言, Gini 策略的题库利用均匀性稍好于 IPA 策略, 且两者均好于 ASI 策略。当在 DINA 模型下属性间高相关时, 4 种选题策略的题库利用率指标值比较接近, 而在其他条件下差异较大。4 种选题策略的题库利用均匀性指标既与 CDM 有关, 又与被试知识状态的分布有关。

图 3 表明, 随测验长度的增加, 各选题策略的卡方值在下降, 即题库使用均匀性逐渐提高。每种选题策略在不同条件下的曲线变化基本相似, JSD 的下降曲线最好, 其次是 Gini 策略, 当在 DINA 模型下属性间高相关时, 4 种选题策略的下降曲线基本重合。

### 5.4　选题用时的比较

表 4 表明, JSD 策略的选题用时最少, 其次是 ASI 策略, 接着是 Gini 策略, 用时最多的 IPA 策略。IPA 策略的选题用时是 Gini 策略的近 10 倍。每种选题策略在不同条件下用时基本不变, 因为选题时间主要与选题策略算法的运算量, 属性个数和题库容量有关, 当属性个数确定和题库容量已知, 选题算法的运算量起决定性作用。

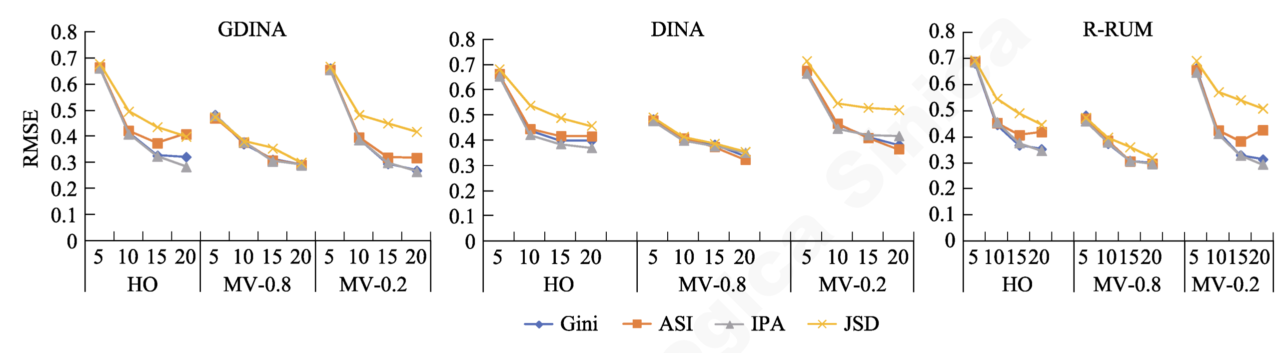
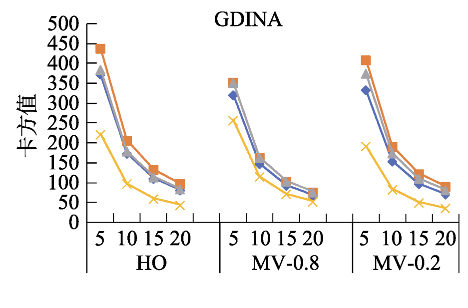
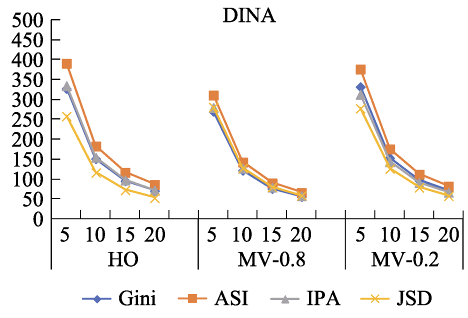



图 2　不同测验长度的能力估计均方差

**表 3　20 题各选题策略的题库使用均匀性指标**

| CDM 模型 | 知识状态生成模型 | 选题策略 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Gini | | ASI | | IPA | | JSD | |
| | | $\chi^2$ | *TOE* | $\chi^2$ | *TOE* | $\chi^2$ | *TOE* | $\chi^2$ | *TOE* |
| G-DINA | HO | 82.38 | 0.41 | 98.75 | 0.47 | 85.34 | 0.42 | **44.45** | **0.26** |
| | MV-0.8 | 69.37 | 0.36 | 77.30 | 0.39 | 77.11 | 0.39 | **53.26** | **0.29** |
| | MV-0.2 | 72.50 | 0.37 | 91.36 | 0.44 | 82.94 | 0.41 | **37.08** | **0.23** |
| DINA | HO | 70.91 | 0.36 | 86.88 | 0.43 | 72.68 | 0.37 | **53.52** | **0.29** |
| | MV-0.8 | **56.55** | **0.31** | 66.74 | 0.35 | 58.98 | 0.32 | 59.31 | 0.32 |
| | MV-0.2 | 72.11 | 0.37 | 83.17 | 0.41 | 67.31 | 0.35 | **58.41** | **0.31** |
| R-RUM | HO | 95.78 | 0.46 | 109.29 | 0.52 | 94.55 | 0.46 | **58.22** | **0.31** |
| | MV-0.8 | 85.70 | 0.42 | 84.99 | 0.42 | 87.92 | 0.43 | **56.27** | **0.30** |
| | MV-0.2 | 88.92 | 0.44 | 105.01 | 0.50 | 95.48 | 0.46 | **60.78** | **0.32** |



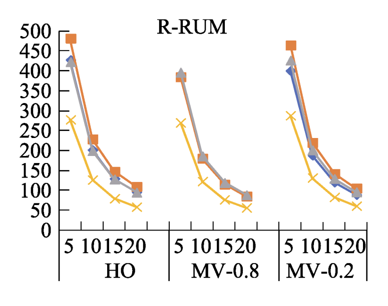




图 3　不同测验长度的卡方值

**表 4 20 题各选题策略的选题用时指标(单位：秒)**

| CDM 模型 | 知识状态生成模型 | 选题策略 | | | |
|---|---|---|---|---|---|
| | | Gini | ASI | IPA | JSD |
| G-DINA | HO | 2.27 | 0.82 | 22.27 | **0.16** |
| | MV-0.8 | 2.27 | 0.82 | 21.95 | **0.16** |
| | MV-0.2 | 2.27 | 0.81 | 22.18 | **0.16** |
| DINA | HO | 2.27 | 0.81 | 21.96 | **0.16** |
| | MV-0.8 | 2.28 | 0.80 | 21.91 | **0.16** |
| | MV-0.2 | 2.26 | 0.78 | 22.04 | **0.16** |
| R-RUM | HO | 2.28 | 0.86 | 21.96 | **0.16** |
| | MV-0.8 | 2.27 | 0.81 | 22.14 | **0.16** |
| | MV-0.2 | 2.26 | 0.81 | 22.01 | **0.16** |

# 6 总结和讨论

## 6.1 总结

本文利用基尼指数的优良性质，构造一种新的双目标 CD-CAT 的选题策略，模拟实验表明新策略的测量精度较高，兼顾题库利用均匀性并能快速实时响应，为同时兼顾宏观能力评估和微观认知诊断提供了新的更优的方法。

实验考察了 3 种 CDM 和 3 种不同被试知识状态分布下，4 种双目标选题策略(Gini 策略、ASI 策略、IPA 策略和 JSD 策略)的表现，综合来看，得到如下结论：(1) Gini 策略和 IPA 策略在分类精度指标，能力估计精度指标和题库使用均匀性指标上均具有相似的表现，测量精度高且受 CDM 模型和被试知识状态分布的影响较小，可以适用于实际测验中含多种认知诊断模型的混合题库。总体而言，Gini 策略稍好于 IPA 策略，且 Gini 策略的选题用时仅为 IPA 策略的十分之一; (2) Gini 策略和 ASI 策略都是两种信息量线性加权的组合策略，在短测验时，两种选题策略在测量精度指标上的表现很接近，而在中长测验时，虽然 ASI 策略的用时是 Gini 策略的 1/3，但 ASI 策略的测量精度和题库使用均匀性均不如 Gini 策略; (3) Gini 策略与 JSD 策略相比，JSD 策略在题库使用均匀性和选题用时指标上有较大的优势，但其测量精度远不如 Gini 策略。

综上所述，短测验时，Gini 策略、IPA 策略和 ASI 策略均有较好的测量精度，都值得推荐。对于中长测验时，对于属性个数少和题库容量较小的情况下，推荐使用 Gini 策略和 IPA 策略，而当属性个数增多和题库容量增大时，推荐使用 Gini 策略。当属性间高相关且属性个数非常多和题库容量非常大时，推荐使用 ASI 策略和 JSD 策略，ASI 策略的测验精度稍高于 JSD 策略。

## 6.2 讨论

Gini 策略是基于被试知识状态类别的后验概率和被试能力估计置信区间的后验概率构造的，因此受 CDM 和被试知识状态分布的影响较小，这种构造方法直接反映后验概率的变化且采用了最小错误率贝叶斯决策确定被试的知识状态，因而测量的精度也非常高。基尼指数的线性加权方式，使得其对后验概率的变化相比熵而言更加敏感，从而有助于扩大选题范围提高题库利用均匀性，且加法运算速度较快，能满足 Dual-CAT 实时响应的需求。

在某些条件下(如被试的知识状态由高阶模型生成), Gini 策略的能力估计精度会稍低于 IPA 策略，而此时 Gini 策略的模式判准率会稍高于 IPA 策略，可能的原因是组合策略中能力的信息量和知识状态的信息量共同作用选择下一题，两种信息量在选题过程中互相均衡的结果。Zheng 和 Chang (2016)指出当已知题库参数，公式(3)中的 KL 信息量可以预先计算，缩短了 ASI 策略的选题用时，而 Gini 策略是定义在随机变量后验概率，必须根据被试的作答反应实时计算，因此选题用时会稍有增加。

JSD 策略仅计算基于当前估计值的 KL 距离，运算量小，选题非常快，而 Gini 策略需考虑有限集合和区间范围内后验概率变化，需要求和与积分运算，因此选题耗时会超过 ASI 策略和 JSD 策略。当测验长度较短时，能力估计值和被试知识状态估计值偏离真值较远，基于他们当前估计值的 JSD 策略的选题范围比较宽泛，从而使得题库的利用率会更加均匀; Gini 策略不依赖于能力和知识状态的当前估计值，而依赖于他们的概率分布，选题会更趋集中。

Gini 策略的测验精较高，但其题库利用率不如 JSD 策略。Wang 等人(2011)的研究表明限制渐进法(Restrictive Progressive Method：RP)和限制阈值法(Restrictive Threshold Method：RT)能均衡测量精度和项目曝光率，下一步研究拟将 Gini 策略与 RP 和 RT 方法结合，提高 Gini 策略的题库利用均匀性。测量精度和题库利用均匀性是一对相互冲突的指标。使用控制项目曝光技术后，题库利用均匀性会更好，但也会带来测量精度下降的不利影响，如何权衡需要进一步研究。另外，使用控制项目曝光技术后，各选题策略之间的差异是否会消除，也有待进一步研究。当属性个数较多时和题库容量较大时，Gini 策略的选题用时可能会超过用户的期望值(延

时超 2 秒) (Nah, 2004), 下一步研究拟将 Gini 策略与动态搜索算法(Zheng & Wang, 2017)结合, 对其优化以减少选题用时。

本文采用分离建模的方法获得两类模型的参数来构建 Dual-CAT 的题库, 题库项目是否完全拟合所关注的模型还需要进一步探查以期获得更准确的测量结果。文中 Dual-CAT 的题库参数的建立过程是先模拟 CDM 的参数和项目的属性向量, 根据 CDM 模型获得反应数据, 然后用反应数据估计 IRT 参数, 这是目前研究中常用的方法(Dai et al., 2016; Kang et al., 2017; Wang et al., 2012, 2014), 能否采用先模拟 IRT 的项目参数, 根据 IRT 模型获得反应数据, 然后用反应数据估计 CDM 参数和项目属性向量的方法构建题库? 在这种方式构建题库下各选题策略的表现有待进一步探查。

随着测验数据的复杂性和测验要求的限定, 选题策略的发展也要适应新测验形式的发展, 比如属性多级化项目测验(涂冬波, 蔡艳, 2015)、多级评分项目测验(蔡艳　等, 2016)、多维项目测验(韩雨婷　等, 2018; Hsu & Wang, 2019)、多阶段 CD-CAT (罗芬　等, 2018; Kaplan & de la Torre, 2020)、融入非统计约束的多阶段测验(Lin & Chang, 2019; Liu et al., 2018)以及结合反应时的 CAT 测验(Fan et al., 2012; Huang, 2020), 可探讨基于基尼指数的选题策略在这些测验场景下的效果及其应用。

## 参　考　文　献


Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431–444.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*, Chapman & Hall / CRC, Boca Raton, FL.

Cai, Y., Miao, Y., & Tu, D. B. (2016). The polytomously scored cognitive diagnosis computerized adaptive testing, *Acta Psychologica Sinica*, *48*(10), 1338–1346.

[蔡艳, 苗莹, 涂冬波. (2016). 多级评分的认知诊断计算机化适应测验. *心理学报*, *48*(10), 1338–1346.]

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the renvironment. *Journal of Statistical Software*, *48*(6), 1–29.

Chang, H. -H., & Ying, Z. L. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*(3), 213–229.

Chen, P., Li, Z., & Xin, T. (2011). A note on the uniformity of item bank usage in cognitive diagnostic computerized adaptive testing. *Studies of Psychology and Behavior*, *37*(1), 212–216.

[陈平, 李珍, 辛涛. (2011). 认知诊断计算机化自适应测验的题库使用均匀性初探. *心理与行为研究*, *37*(1), 212–216.]

Cheng, Y. (2007). *The dual information method for item selection in cognitive diagnostic computerized adaptive testing* (Unpublished Master's thesis). University of Illinois at Urbana–Champaign.

Cheng Y. (2009). When cognitive diagnosis meets computerized adaptive testing. *Psychometrika*. *74*(4), 619–632.

Cheng, Y., & Chang, H. -H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *62*(2), 369–383.

Dai, B. Y., Zhang, M. Q., & Li, G. M. (2016). Exploration of item selection in dual purpose cognitive diagnostic computerized adaptive testing: Based on the RRUM. *Applied Psychological Measurement*, *40*(8), 625–640.

Du, X. X. (2010). *A new strategy of item selection of cognitive diagnosis computerized adaptive testing* (Unpublished Master's thesis). Jiangxi Normal University, Nanchang, China.

[杜宣宣. (2010). *具有认知诊断功能的计算机化自适应测验的选题策略研究*(硕士学位论文). 江西师范大学, 南昌.]

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333–353.

Fan, Z. W., Wang, C., Chang, H. -H., & Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics*, *37*(5), 655–670.

Han, Y. T., Gao, X. L., Wang, D. X., Cai, Y., & Tu, D. B. (2018). Item selection methods in multidimensional polytomous computerized adaptive testing. *Journal of Psychological Science*, *41*(6), 1500–1507.

[韩雨婷, 高旭亮, 汪大勋, 蔡艳, 涂冬波. (2018). 多级评分项目的多维 CAT 选题策略开发. *心理科学*, *41*(6), 1500–1507.]

Hartz, S. M. (2002). *A bayesian framework for the unified model for assessing cognitive abilities: blending theory with practicality* (Unpublished Doctoral dissertation). University of Illinois at Urbana–Champaign, Urbana-Champaign, IL.

Hsu, C. -L., & Wang, W. -C. (2015). Variable-length computerized adaptive testing using the higher order DINA model. *Journal of Educational Measurement*, *52*(2), 125–143.

Hsu, C. -L., & Wang, W. -C. (2019). Multidimensional computerized adaptive testing using non-compensatory item response theory models. *Applied Psychological Measurement*, *43*(6), 464–480.

Huang, H. -Y. (2020). Utilizing response times in cognitive diagnostic computerized adaptive testing under the higher-order deterministic input, noisy 'and' gate model. *British Journal of Mathematical and Statistical Psychology, 73(1),* 109–141.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258–272.

Kang, H. -A., Zhang, S. S., & Chang, H. -H. (2017). Dual-objective item selection criteria in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement*, *54*(2), 165–183.

Kaplan, M., & de la Torre, J. (2020). A blocked-CAT procedure for CD-CAT. *Applied Psychological Measurement*, *44*(1), 49–64.

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement*,

39(3), 167–188.
Li, H. (2012). *Statistical learning method*. Beijing: Tsinghua University Press.
[李航. (2012). *统计学习方法*. 北京: 清华大学出版社.]
Lin, C. -J., & Chang, H. -H. (2019). Item selection criteria with practical constraints in cognitive diagnostic computerized adaptive testing. *Educational and Psychological Measurement*, *79*(2), 335–357.
Liu, S. C., Cai, Y., & Tu, D. B. (2018). On-the-fly constraint-controlled assembly methods for multistage adaptive testing for cognitive diagnosis. *Journal of Educational Measurement*, *55*(4), 595–613.
Lord, M. F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.
Luo, F., Wang, X. Q., Ding, S. L., & Xiong, J. H. (2018). The design and selection strategies of adaptive multigroup Testing for Cognitive Diagnosis. *Journal of Psychological Science*, *41*(3), 720–726.
[罗芬, 王晓庆, 丁树良, 熊建华. (2018). 自适应分组认知诊断测验设计及其选题策略. *心理科学*, *41*(3), 720–726.]
Ma, W. C., & de la Torre, J. (2020). *GDINA: The generalized DINA model framework*. R package version 2.7.9, https://CRAN.R-project.org/package=GDINA.
McGlohen, M. K., & Chang, H. -H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods*, *40*(3), 808–821.
Nah, F. F. -H. (2004). A study on tolerable waiting time: How long are web users willing to wait? *Behaviour and Information Technology*, *23*(3), 153–163.
Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106.
Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann, San Mateo, CA.
Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: theory, method, and application*. New York: The Guilford Press.
Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, *51*(3), 337–350.
Tu, D. B., & Cai, Y. (2015). The Development of CD-CAT with polytomous attributes. *Acta Psychologica Sinica*, *47*(11), 1405–1414.
[涂冬波, 蔡艳. (2015). 基于属性多级化的认知诊断计算机化自适应测验设计与实现. *心理学报*, *47*(11), 1405–1414.]
Veerkamp, W. J. J., & Berger, M. P. F. (1994). *Some new item selection criteria for adaptive testing* (Research Rep. 94-6). Enschede, The Netherlands: University of Twente, Department of Educational Measurement and Data Analysis.
Wang, C., & Chang, H. -H. (2011). Item selection in multidimensional computerized adaptive testing-gaining information from different angles. *Psychometrika*, *76*(3), 363–384.
Wang, C., Chang, H. -H., & Douglas, J. (2012). Combining CAT with cognitive diagnosis: A weighted item selection approach. *Behavior Research Methods*, *44*(1), 95–109.
Wang, C., Chang, H. -H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement*, *48*(3), 255–273.
Wang, C., Zheng, C. J., & Chang, H. -H. (2014). An enhanced approach to combine item response theory with cognitive diagnosis in adaptive testing. *Journal of Educational Measurement*, *51*(4), 358–380.
Xu, X. L., Chang, H. -H., & Douglas, J. (2003, April). *A simulation study to compare CAT strategies for cognitive diagnosis*. Paper presented at the annual meeting of National Council on Measurement in Education, Chicago, IL.
Zhang, X. G. (2010). *Pattern recognition (Third Edition)*. Beijing: Tsinghua University Press.
[张学工. (2010). *模式识别(第三版)*. 北京: 清华大学出版社.]
Zheng, C. J., & Chang, H. -H. (2016). High-efficiency response distribution–based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement*, *40*(8), 608–624.
Zheng, C. J., He, G., & Gao, C. L. (2018). The information product methods: A unified approach to dual-purpose computerized adaptive testing. *Applied Psychological Measurement*, *42*(4), 321–324.
Zheng, C. J., & Wang, C. (2017). Application of binary searching for item exposure control in cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement*, *41*(7), 561–576.
Zhou, Z. H. (2016). *Machine learning*. Beijing: Tsinghua University Press.
[周志华. (2016). *机器学习*. 北京：清华大学出版社.]

## A new dual-objective CD-CAT item selection method based on the Gini index

LUO Fen[1,2], WANG Xiaoqing[2], CAI Yan[1], TU Dongbo[1]

([1] *School of Psychology, Jiangxi Normal University, Nanchang* 330022, *China*)

([2] *College of Computer Information Engineering, Jiangxi Normal University, Nanchang* 330022, *China*)

**Abstract**

Existing literature has shown that dual-objective CD-CAT testing can facilitate the achievement of measurement objectives for both formative and summative assessments. And the Gini Index can be used as a measurement for the degree of uncertainty of random variables since a smaller Gini value indicates a lower degree of uncertainty. Hence, this paper proposed a Gini-Index-based selection method for dual-objective CD-CAT, and it measured the changes in the posterior probability of knowledge state and confidence interval for

latent traits estimation. By adopting the Bayesian Decision Theory, the potential information of participants could be detected based on participants' responses and changes in posterior probability distribution of two the random variables.

Monte Carlo Simulation was used to test the performances of the selection method based on Gini, ASI, IPA and JSD, respectively. The item banks measured 5 attributes consisting of 250 items in total, and each item measured 3 attributes at most. The true knowledge state of each participant was generated by HO-CDM and Multivariate Normal Models (both means were 0 and covariance coefficient was 0.8 and 0.2, respectively). G-DINA, DINA and R-RUM were adopted as the cognitive diagnostic models and the item bank of each of these three models included both CDM and 2PL parameters. Specifically, CDM parameters were generated by a G-DINA package in R software with the slipping and guessing parameters randomly selected from uniform distribution in a range from 0.05 to 0.25. The 2PL parameters were estimated by factoring in the responses elicited from 3, 000 participants' responses to all items in item banks using the mirt package. Four indexes, namely the pattern match ratio, root mean square error of latent trait, chi-square value and time needed for item selection, were adopted in comparing the efficiency of different item selection methods. The value for each index was the mean of 10 repeated simulations of 1, 000 participants' responses to all item bank.

The results showed that (1) The Gini and IPA selection methods had similar performance in terms of pattern match ratio, root mean square error of latent trait and chi-square value. Both methods were high in precision measurement and low in sensitivity to CDM and the distribution of participants' cognitive patterns, making both methods applicable to the item banks featuring a mixture of cognitive diagnosis models. By comparison, the Gini method outperformed slightly the IPA method in pattern match ratio and time needed for item selection in which the Gini method was only one-tenth that of the IPA method; (2) Both the Gini and ASI selection methods were weighted linear combination approaches. The performances of the two methods were very close in the short test. In the long test, however, although time needed for item selection using the ASI method was only one-third that of the Gini method, the latter was superior to the former in terms of measurement accuracy and chi-square value; (3) Although the JSD method outperformed the Gini method in terms of uniformity of item bank usage and time needed for item selection, its measurement accuracy was far less than the latter.

To summarize, the Gini, IPA and ASI selection methods all have good measurement accuracy and hence are all recommended for short tests. For medium and long tests with a limited number of attributes and a smaller item bank, the Gini and IPA selection methods are recommended. As the number of attributes and item bank size grow, the Gini method is recommended. When there are high correlations among different attributes, as well as a large number of attributes and big item bank size, the ASI and JSD selection methods are recommended with the ASI method slightly outperforming the JSD method in measurement accuracy.

**Key words**　cognitive diagnostic, items response theory, Gini index, dual objective CD-CAT, selection method